COV-19 Infection Rate, Masking Mandate and Vaccine Rate Data Analysis

Weiqing (Ivy) Lin

Department of Data Science, University of Washington

DATA 512: Human-Centered Data Science

Introduction

This is our final project of Data512 Human-Centered Data Science. Our goal is to practice how to complete a whole data science research project, from gathering data, processing data, setting up our researching questions (Null Hypothesis), proposing the proper methodologies, then analyzing with R/Python or other tools, finally leading to the conclusion, and summarizing with a constructed research report.

Since Cov-19 pandemic, everyone's life changed. Most of us are now remoting from work or taking online classes for schools. People like me, who are so "foodie", really want to go checking their favorite restaurants. But with the government rules, people cannot enter any restaurants without vaccine cards. There are not a few people very anti-vaccine, people think this is their freedom to choose if they take vaccine shot or not. Even though the government is trying different rewards or pushing strict rules, the situation didn't change a lot. We can see the vaccine rate was rising quickly at the begging, since people like me, who trust the vaccine benefits, are eager to take vaccine to protect themselves and their families. But after the vaccine rate reached a contain point, no more increasing, because the rest of people are not willing to take vaccine based on their different situations. Same as adding WA Notify to your smartphone to alert you if you may have been exposed to COVID-19, and anonymously alert others if you test positive. Some people like this idea, but a lot of people have more privacy concern to be tracked.

With Delta Variant hit us hard, "Back to Office" date has been pushed again and again based on the high infection rates. Meanwhile, new Omicron has turned out to be highly transmissible and less susceptible to vaccines than other variants. CDC guidance for Cov-19 vaccine doses and booster shot are encouraged everywhere on medias. So does Cov-19 vaccine really help? Maybe it doesn't help a lot for Delta and other variants.

In order to answering these questions, I used Cov-19 related data around Daily Cases, Infection Rate and Vaccine Data, and tried to find the correlation between 7 days rolling average Infection Rate and Vaccine Rate. All analysis are performed in a single Jupiter notebook data-512-final-project/hcds-final-project.ipynb at main · IvyLinMS/data-512-final-project (github.com).

Background/Related Work

Around the topic of taking vaccine or not, there're a lot of discussions. Most of them are personal feeling in blogs, newspapers, and other social medias. Robert Hart, Forbes Staff wrote one interesting article By The Numbers: Who's Refusing Covid Vaccinations—And Why (forbes.com), he listed all the numbers of people unwilling or refusing to vaccinations and tried to seek Why, but this article doesn't really apply any data analysis itself, it just based on the polling or some survey data from Kaiser Family Foundation.

There is another good article on Infected, Vaccinated, or Both: How Protected Am I From COVID? (webmd.com) by Brenda Goodman, MA. The author mentioned "According to CDC data, at the height of the Delta surge in August, fully vaccinated people were six times less likely to get a COVID-19 infection compared with unvaccinated people, and 11 times less likely to die if they did get it." I think this is very encouraging for everyone to take vaccine. The author also linked us to another post by Scott Gottlieb, MD

on Twitter about 70% of population vaccinated potentially keep death from Covid down. In Figure1, the chart "Cumulative deaths attributed to Cov-19 in US, UK and South Korea" shows us the different slopes of Cumulative deaths between US, UK and South Korea, which has the 70% population vaccinated.

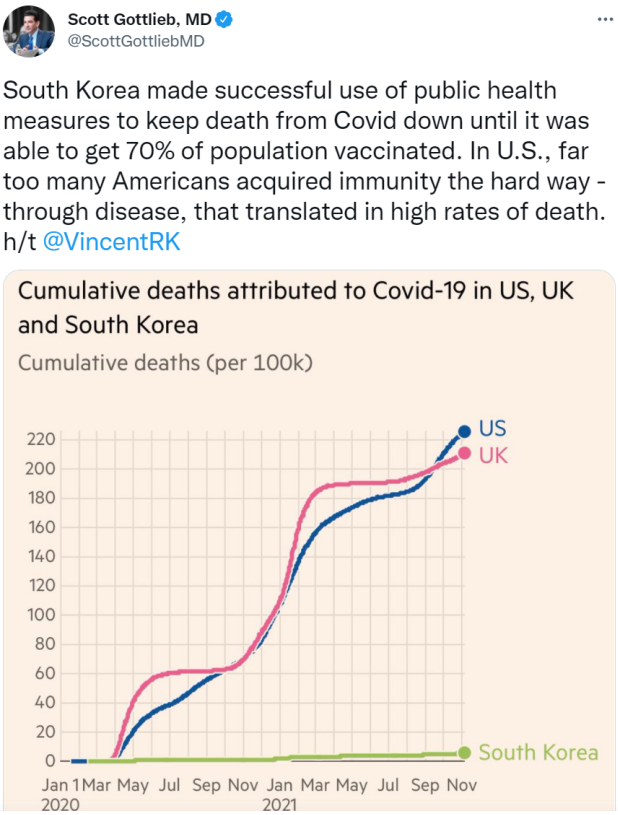

Figure 1

In addition, there's another peer-reviewed publication around Vaccination Rates with Cases Numbers, called Relationship Between COVID-19 Cases and Vaccination Rates in New York State Counties (stfm.org), which was published on PRiMER. 2021;5:35. The authors calculated descriptive statistics and bivariate correlations for vaccination rates and cases across counties in New York State. Later, they used linear regression using cases/100K population per NYS county, frozen at a single snapshot in time, as the outcome variable, predicted by percentage of each county's population (completed series/two doses), controlling for county population. It's very close to my project goal to find the correlation between 7 days rolling average Infection Rate and Vaccine Rate, but not exactly same as my Null Hypothesis question "There is no correlation between Vaccine Rate and Daily Infection Rate of 7 days rolling average".

Methodology

For this whole Cov-19 research project, first we started with A4 common analysis, then it's A5 extension plan. In A4 assignment, I used the RAW_us_confirmed_cases.csv file from the Kaggle repository of John Hopkins University COVID-19 data, the CDC dataset of masking mandates by county, and New York Times mask compliance survey data. What I did was some basic exploratory data analysis in a single

Jupiter notebook, there were no fancy statistical methods/models involved. I chose to mainly use data visualizations to present my results. One thing I did after data cleaning is to standardize among the three datasets. Since **FIPS** is the key column to join the data, but it has different formatting everywhere, RAW_us_confirmed_cases dataset has a float data type of FIPS, The CDC dataset of masking mandates by county has two separate columns FIPS_State and FIPS_County. Besides, pivot data is another method I used a lot in data processing. I did pivot(pd.melt) the Date column for RAW_us_confirmed_cases dataset, also the Response and Proportion column for The New York Times mask compliance survey dataset.
e.g.:

```
df_mask_use_palm_beach_transformed =pd.melt(df_mask_use_by_county_transformed, var_name ='Response', value_name = "Proportion")
```

What's more, to solve the problem when some of my data is not available for further research, since my assigned County Palm Beach, FL has all "NaN" values for Face_Masks_Required_in_Public column in the CDC dataset of masking mandates by county. To handle this issue, I brought in the additional data for comparison (another county Spotsylvania, VA, which has the similar Prevalence of Mask Wearing around ALWAYS = 0.784). With the new data added in, we're able to observe that the mandatory mask policy helped a lot to prevent the virus spreading, and even after the mandatory period, it continued to keep the Infection Rate lower since people are used to wearing masks in public areas.

For A5 extension plan, my focus is to find the correlation between 7 days rolling average Infection Rate and Vaccine Rate. The Additional data I brought in is the CDC dataset of Reporting County-Level COVID-19 Vaccination Data | CDC. First of all, I did the exploring data visualization to find any trend and relationship between Infection Rate and Vaccine Rate. Then I noticed between 2021-01 and 2021-07, the Infection Rate dropped dramatically when the Vaccine Rate raised in Palm Beach. Then I conducted the statistical analysis by using Panda .corr() function to find the correlation between Infections Rate and Vaccine Rate. Considering Delta Variant's Impact, I split the data to 3 time periods, 1st period is before 2021-01 Vaccine started, 2nd period is after 2021-01 and before 2021-07 Delta Variant, then 3rd period is after 2021-07 Delta variant came into our picture. By visualizing via seaborn heatmap, we can see there's no correlation between Infection Rate and Vaccine Rate before 2021-01 Vaccine started, then very strong negative correlation between Infection Rate and Vaccine Rate after 2021-01 Vaccine started and before 2021-07 Delta Variant, meaning Cov-19 Vaccine helped a lot. After 2021-07 Delta variant started, there's very slight negative correlation between Infection Rate and Vaccine Rate, meaning Vaccine didn't help a lot for stopping Delta variant spreading.

Furthermore, I also used Linear Regression to build a model to find the coefficient between the Vaccine rate and 7 days rolling average of daily infection rate. The model's coefficient is -0.00088422, which is very small but not zero and is negative, which indicates that when Vaccine rate increase, 7 days rolling average of daily infection rate will drop. To further confirm if the coefficient is correct, I set a NULL Hypothesis that there is no correlation between vaccine rate and 7 days rolling average of infection rate, and then used the statsmodels.api OLS Regression Results to find out what's the P_Value for this Null Hypothesis. Besides, we can also look into other values of Ordinary Least Squares Regression Results. For example, R-squared and F-statistic are other signs for us to tells the goodness of fit of a Regression.

Findings

My GitHub repo for your reference: [IvyLinMS/data-512-final-project (github.com)]

First of all, I did one chart (Figure 2) shows Daily New Covid Cases by 7 days on average. Y-axis is the number of Covid cases. The blue line in the chart shows the Daily New Covid Cases, the orange line shows the 7 days rolling average, which means every day we look back 7 days to get the average. Why 7 days rolling average? Because we can avoid the impact of every Monday reporting case spike and Cov-19 incubation period. We can see clearly the Spikes of Daily New Covid Cases are 2020-08 summer break, and holiday seasons between 2020-11 and 2021-01, then 2021-08 and 2021-09 summer break.

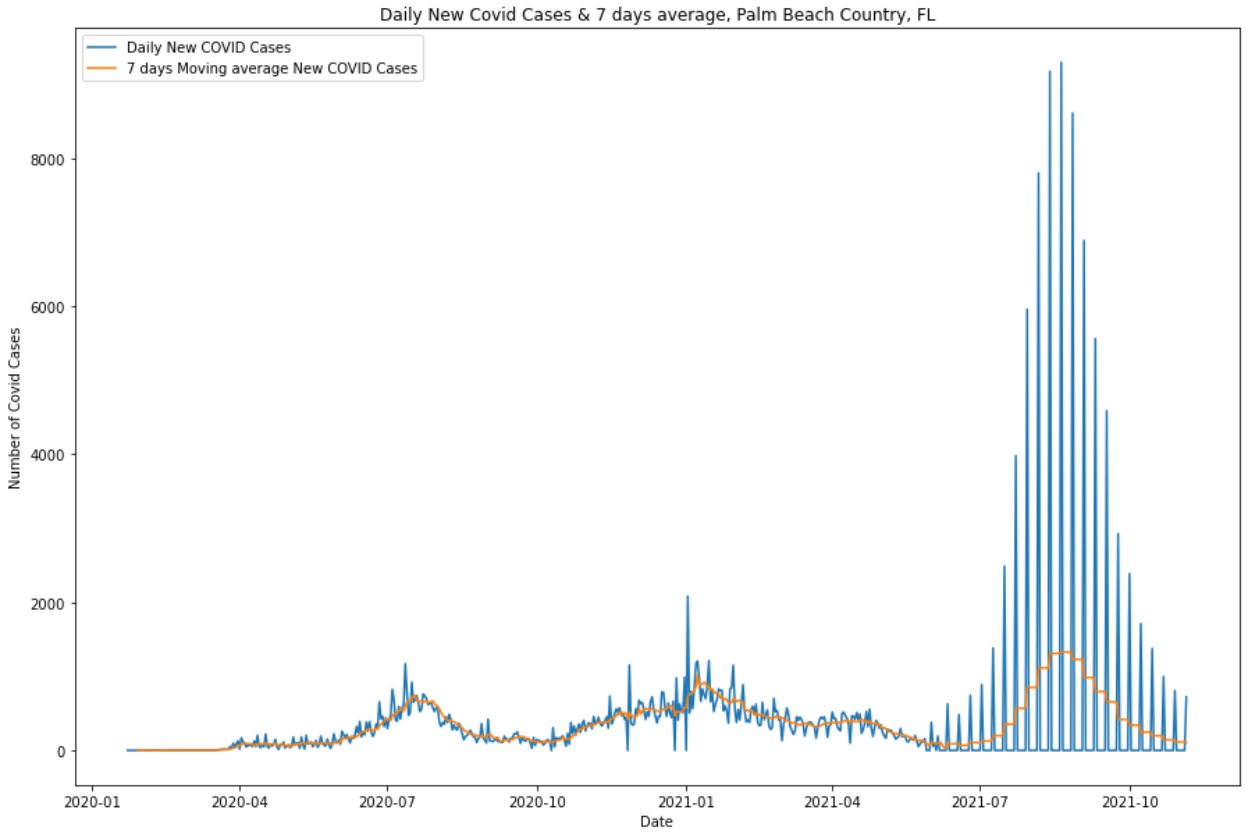

Figure 2

Then instead of case numbers, I did two charts for average infection rate. A-axis shows the time series. In Figure 3, Y-axis shows 7 days rolling average infection rate. Why I prefer Infection Rate not Cov-19 case numbers? Since Infection Rate is daily new case / population, which we can reduce the impact of different populations. As we can see, the infection rate also indicates the spikes of summer breaks and holiday season. This is as expected, since the Palm Beach area in Florida is a tourism city, so both the Daily New Covid Cases and infection rate went higher during travel seasons. What surprised me is that both the case numbers and infection rate of the year 2021 summer break are higher than the year 2020, when most people were vaccinated. I wonder if it's related to the more contagious Delta variant. In Figure 4, other

than the same spikes of summer breaks and holiday season, we can also notice the Infection Rate diff went negative after the 2021 summer break, this might be a good sign of the declining Delta variant.

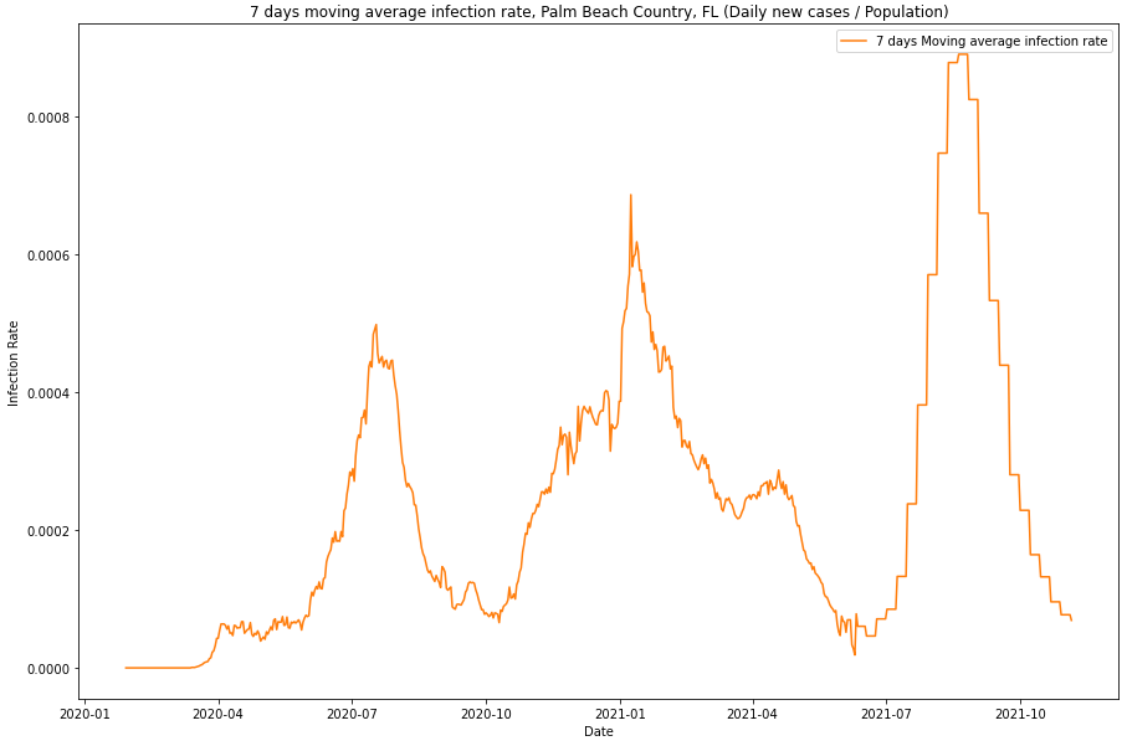

Figure 3

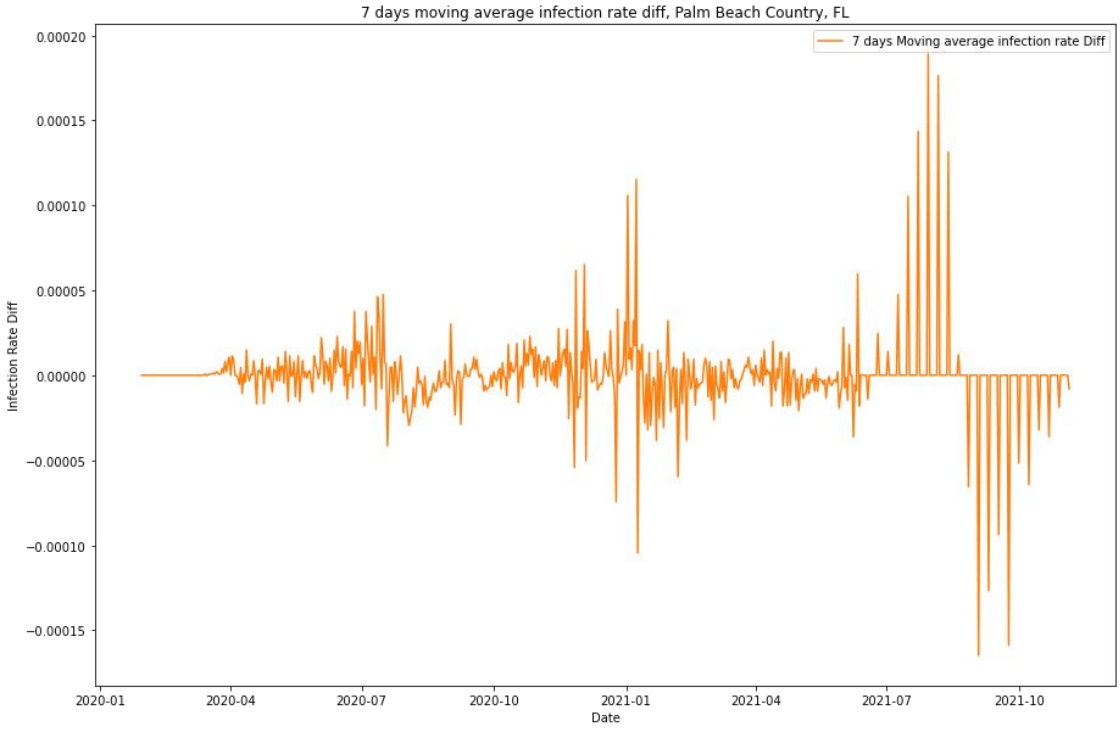

Figure 4

Furthermore, I tried to analyze the masking policies' impact. But since Palm Beach, FL doesn't have a mandatory mask policy, I had to find another county Spotsylvania, VA, which has the similar Prevalence of Mask Wearing around 78.4% population for ALWAYS wearing mask.

	COUNTYFP	NEVER	RARELY	SOMETIMES	FREQUENTLY	ALWAYS
0	12099	0.03	0.02	0.05	0.116	0.784

In Figure 5, I did a chart of 7 days rolling average infection rate comparing Palm Beach, FL with Spotsylvania, VA. The blue line is for Palm Beach, FL, the orange line is for Spotsylvania, VA. The green area is the period when people were required wearing masks mandatorily in Spotsylvania, VA. Besides the higher infection rate during summer breaks and holiday seasons, we can also see the mandatory mask policy helped a lot when it began in the green box. The orange line shows Spotsylvania's infection rate was way lower than the blue line for Palm Beach's. Even after the green period of mandatory mask policy, people got used to wearing masks in Spotsylvania, which kept the infection rate continuing lower than Palm Beach, where masking was never mandated.

Figure 5

For A5 extension research, my focus is on the correlation between 7 days rolling average Infection Rate and Vaccine Rate. In Figure 6, I did a chart for "7 days rolling average infection rate and Vaccine rate at

Palm Beach, FL", Blue line shows the 7 days rolling average infection rate, Orange line shows the vaccine rate by time series. As we can see, vaccine rate was 0% before 2021-01, then Infection Rate dropped dramatically when the Vaccine Rate raised in Palm Beach between 2021-01 and 2021-07. Later, when Delta Variant started around 2021-07 https://www.newsweek.com/first-us-covid-delta-variant-cases-how-did-it-mutate-1617871, also the impact of government nationwide reopening and summer break's impact, the Infection Rate went up again even though the Vaccine Rate kept raising from around 48% to 60%.

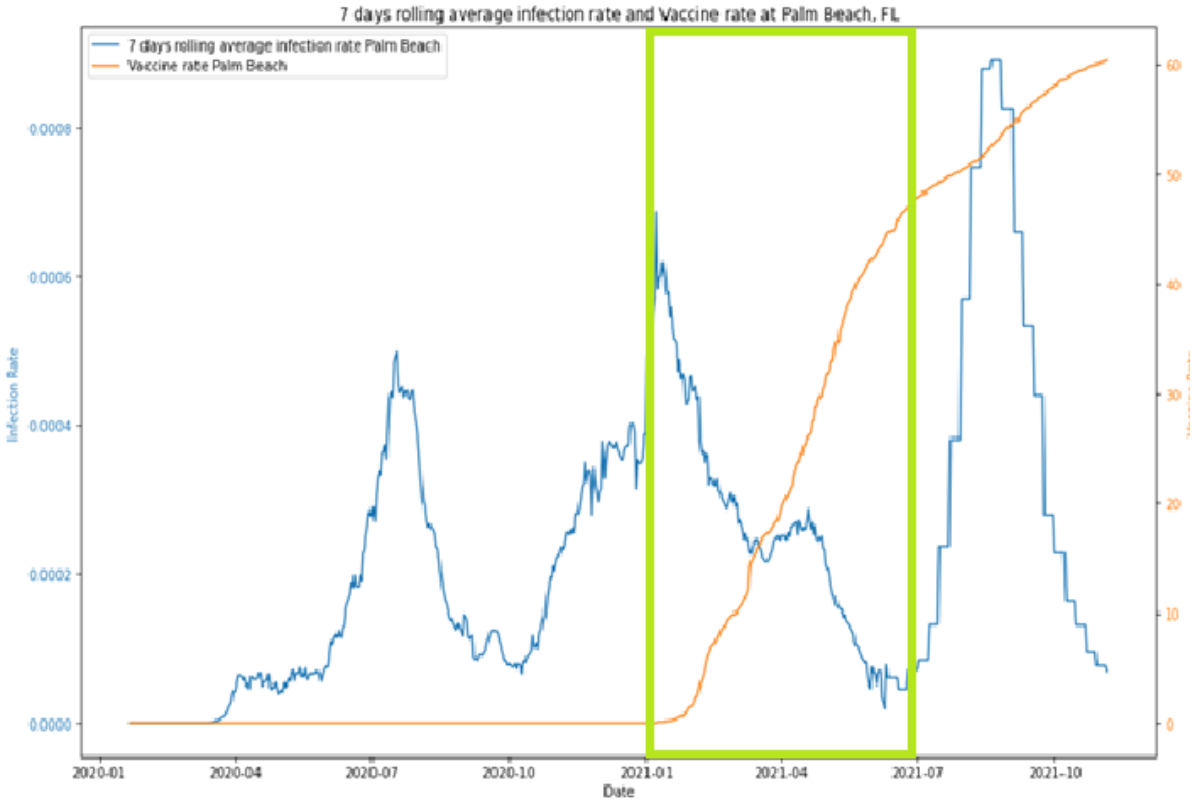

Figure 6

Furthermore, I conducted the statistical analysis by using Pandas .corr() function to find the correlation between Infections Rate and Vaccine Rate. Considering Delta Variant's Impact, I split the data to 3 time periods for analysis:

- Before 2021-01 Vaccine started
- After 2021-01 and before 2021-07 Delta Variant
- After 2021-07 Delta variant came into our picture

In Figure 7, we can see there's no correlation between Infection Rate and Vaccine Rate before 2021-01 Vaccine started. This is as expected, since the vaccine rate was 0% before 2021-01, no vaccine available yet, so the heatmap is showing NaN for no correlation.

	daily_infection_rate_new_cases_moving_average_7_days	Series_Complete_Pop_Pct
daily_infection_rate_new_cases_moving_average_7_days	1.0	NaN
Series_Complete_Pop_Pct	NaN	NaN

```
# Use heatmap to visualize
sns.heatmap(correlations_before_vacacine)
plt.show()

# as expected, there is no correclation
```

Figure 7

In Figure 8, as we can see, the correlation between 7 days rolling average Infection Rate and Vaccine Rate is -0.922376, meaning very strong negative correlation after 2021-01Vaccine started and before 2021-07 Delta Variant. This makes a lot of sense, since the first batch of vaccines were distributed to Seniors and Front workers, then it was gradually available to the public, and now even open to young kids between 5 years to 11 years. That means at the beginning, the limited vaccination was distributed to the most likely infection group of people, which could help to reduce the Case numbers and infection rate sharply. After that, more and more people took the vaccine 1st dose and 2nd dose, the infection rate went down lower and lower till Delta Variant and other fact impacted.

	daily_infection_rate_new_cases_moving_average_7_days	Series_Complete_Pop_Pct
daily_infection_rate_new_cases_moving_average_7_days	1.000000	-0.922376
Series_Complete_Pop_Pct	-0.922376	1.000000

```
sns.heatmap(correlations_after_vacacine)
plt.show()

# we can see there's very strong negative correlation between Infection Rate and Vaccine Rate
```

Figure 8

In Figure 9, we can see the correlation between 7 days rolling average Infection Rate and Vaccine Rate is -0.29999, meaning slightly negative correlation after 2021-07 Delta Variant came into the picture. The other facts like the summer break and nationwide reopening to saving the US economies, the RAW Daily Covid Cases data and the 7 days rolling average Infection Rate can be impacted by travelling, family gathering and public places like restaurants and music concerts reopening. Then the school year started, the Case numbers and Infection Rate can be also impacted by teachers and students in-person learning exposure.

```
sns.heatmap(correlations_after_delta)
plt.show()

# we can see very slight negative correlation between Infection Rate and Vacine Rate after Delta variant
```

Figure 9

On top of this, I also used Linear Regression to build a model to find the coefficient between the Vaccine rate and 7 days moving average of daily infection rate. In the below code snippet, we can see the model's coefficient is -0.00088422, which is very small but not zero and is negative, which indicates that when Vaccine rate increase, 7 days rolling average of daily infection rate will drop.

```
# Train a linear regression model using the vaccine rate and infection rate and get the model's coefficient
model = LinearRegression().fit(vaccine_rate, infection_rate)
model.coef_
```

```
array([-0.00088422])
```

To further confirm if the coefficient is correct, I set a NULL Hypothesis that there is no correlation between Vaccine rate and 7 days rolling average of Infection rate, and then used the statsmodels.api OLS Regression Results to find out what's the P_Value for this Null Hypothesis. Besides, we can also look into other values of OLS Regression Results. For example, R-squared and F-statistic are other signs for us to tell the goodness of fit of a Regression.

Based on the Ordinary Least Squares Regression Results in Figure 10:

- P_Value is approximately 0.000, which is smaller than 0.05, so we can reject NULL Hypothesis: There is no correlation between Vaccine Rate and Daily Infection Rate of 7 days rolling average.
- Coef = -0.0009, it confirms that when Vaccine rate increase, 7 days rolling average of daily infection rate will drop.
- R-squared = 0.851 and very large F-statistic = 1009, telling the goodness of fit of a Regression.

```
# Use Ordinary Least Squares by Call SM.OLS to get p_value
x = vaccine_rate
y = infection_rate
x2 = sm.add_constant(x)
est = sm.OLS(y, x2)
model = est.fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.851
Model:                            OLS   Adj. R-squared:                  0.850
Method:                 Least Squares   F-statistic:                     1009.
Date:                Sun, 12 Dec 2021   Prob (F-statistic):           4.97e-75
Time:                        21:56:25   Log-Likelihood:                 662.16
No. Observations:                 179   AIC:                            -1320.
Df Residuals:                     177   BIC:                            -1314.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0451      0.001     59.901      0.000       0.044       0.047
x1            -0.0009   2.78e-05    -31.767      0.000      -0.001      -0.001
==============================================================================
Omnibus:                       14.528   Durbin-Watson:                   0.088
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               16.479
Skew:                           0.606   Prob(JB):                     0.000264
Kurtosis:                       3.859   Cond. No.                         45.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 10

Discussion/Implications

As mentioned earlier in my introduction of this project, people think this is their freedom to be tracked or not tracked by WA Notify App, or the government should not push people to take Cov-19 vaccine shots. Some people don't like the whole masking mandated requirement. The survey data we got from the New York Times mask compliance data https://github.com/nytimes/covid-19-data/tree/master/mask-use, it can be not 100% accurate. Based on we learnt from this human-centered DS course, sometimes survey itself could be biased and sometimes survey results don't represent the general population, since some people maybe not be willing to take survey. Maybe only half of people, being a good citizen, they are willing to follow government masking rules and then they take surveys, which might result the survey masking % higher than real situation.

Although the data may contain some bias, but the important findings in correlation between Vaccine rate and 7 days moving average of daily infection rate shows us that Cov-19 vaccine helps to control the rising of infection rate. Now is close to holiday seasons, the Christmas family re-unions and friend gathering parties will probably lead toward another wave of Covid cases peak based on history data. I will suggest people, who're unwilling to take vaccine, to re-consider taking Cov-19 doses. For myself, I will take a booster shot before heading to any party, which potentially get me exposed to someone with Covid.

For future research, based on the good fit of Linear Regression model during finding the coefficient between the Vaccine rate and Infection rate, I think we can further build a model to help us predict, when given X% vaccine rate, what infection rate would be for nationwide. Even outside of US, if we could get vaccinate data, daily cases, and infection data from other countries, we can also predict the trend for a period of time and a specific area. But the result might also be impacted by other facts, which is leading us to think about our limitation.

For example, in US we have to consider the summer break, winter break and nationwide reopening, the RAW Daily Covid Cases data can be impacted by travelling, restaurant and music concerts reopening, etc. After winter break and holiday seasons, kids are back to the school, the Case numbers and Infection Rate can be impacted by teachers and students in-person learning exposure. Then Delta and other new variant like Omicron or unknown ones, might also impact a lot for the real situation.

Limitations

What's more, there're more data limitations we need to consider as below:

- For CDC vaccine dataset from Reporting County-Level COVID-19 Vaccination Data | CDC

  CDC's dose number estimates might differ from those reported by jurisdictions and federal entities. People receiving doses are attributed to the jurisdiction in which the person resides. When the vaccine manufacturer is not reported, the recipient is considered fully vaccinated with two doses.

- For New York Times covid-19-data/mask-use at master · nytimes/covid-19-data (github.com)

  My assigned county Palm Beach, FL never has Masking Mandated requirement, which pushed me searching for another similar County by survey data. I found Spotsylvania County, VA has the same prevalence of Mask Wearing around 78.4% population for ALWAYS wearing mask. But actually, Spotsylvania only has 1/10 of population of Palm Beach, FL.

Conclusion

Using the time series data visualization of Palm Beach Accumulate Covid Cases and Infection Rate:

- Total Accumulate Covid Cases spikes happened on summer breaks and holiday seasons
- 7 days rolling average infection rate also showed clearly spikes of 2020 and 2021 summer breaks, and holiday seasons between 2020-11 and 2021-01.
- Using masking mandated County Spotsylvania, VA as comparison, masking policy helped a lot when it began, also kept the infection rate lower than Palm Beach, even after the time period of mandatory mask policy, maybe because people got used to wearing masks.

Using Pandas .corr() function and Seaborn Heatmap, we can see:

- No correlation between Infection Rate and Vaccine Rate before 2021-01 Vaccine started.
- Very Strong Negative correlation between Infection Rate and Vaccine Rate after 2021-01 Vaccine started and before 2021-07 Delta Variant, Vaccine helped a lot.
- Very slight negative correlation between Infection Rate and Vaccine Rate after 2021-07 Delta variant, Vaccine didn't help a lot for Delta variant.

From the statsmodels.api OLS Regression Results:

- P_Value is approximately 0.000, which is smaller than 0.05, so we can reject NULL Hypothesis: There is no correlation between Vaccine Rate and Daily Infection Rate of 7 days rolling average.
- Coef = -0.0009, it confirms that when Vaccine rate increase, 7 days rolling average of daily infection rate will drop.
- R-squared = 0.851 and very large F-statistic = 1009, telling the goodness of fit of a Regression.

References

[1] Halle Cerio, BS | Laura A. Schad, MPH | Telisa M. Stewart, DrPH | Christopher P. Morley, PhD

Relationship Between COVID-19 Cases and Vaccination Rates in New York State Counties (stfm.org)

Published: 9/29/2021 | DOI: 10.22454/PRiMER.2021.432215

[2] Robert Hart, Forbes Staff

By The Numbers: Who's Refusing Covid Vaccinations—And Why (forbes.com)

[3] Brenda Goodman, MA

Infected, Vaccinated, or Both: How Protected Am I From COVID? (webmd.com)

.

Data Sources

- The CDC dataset of masking mandates by county. https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i
    - Licesnse: Public Use (User Agreement: https://www.cdc.gov/nchs/data_access/restrictions.htm)

- The RAW_us_confirmed_cases.csv file from the Kaggle repository of John Hopkins University COVID-19 data. https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university?select=RAW_us_confirmed_cases.csv
    - License: Attribution 4.0 International (CC BY 4.0)
- The New York Times mask compliance survey data. https://github.com/nytimes/covid-19-data/tree/master/mask-use
    - License: Copyright 2021 by The New York Times Company
- CDC's COVID Data Tracker provides COVID-19 vaccination data in the United States https://www.cdc.gov/coronavirus/2019-ncov/vaccines/distributing/reporting-counties.html
    - Licesnse: Public Domain U.S. Government
    - Overall US COVID-19 Vaccine administration and vaccine equity data at county level.
    - The dataset has 1.09M Rows and 32 Columns